基于特征选择与集成学习的钓鱼网站检测方法

周传华[1,2]，柳智才[1?]，丁敬安[1]，周家亿[3]

(1. 安徽工业大学 管理科学与工程学院，安徽 马鞍山 243002; 2. 中国科学技术大学 计算机科学与技术学院，合肥 230026; 3. 早稻田大学 IPS 学院，日本 东京)

摘　要：针对目前大部分钓鱼网站检测方法存在检测准确率低、误判率高等问题，提出了一种基于特征选择与集成学习的钓鱼网站检测方法。首先使用 FSIGR 算法进行特征选择，该算法结合过滤和封装模式的优点，从信息相关性和分类能力两个方面对特征进行综合度量，并采用前向递增后向递归剔除策略对特征进行选择，以分类精度作为评价指标对特征子集进行评价与选择，从而获取最优特征子集；然后使用选择后的最优特征子集基于随机森林集成学习分类算法进行训练。在 UCI 数据集上的实验表明，所提方法能够有效提高钓鱼网站检测的正确率，降低误判率，具有实际应用意义。

关键词：钓鱼网站；随机森林；信息增益率；特征选择

中图分类号：TP393.08　　doi: 10.3969/j.issn.1001-3695.2017.10.0998

Method of phishing website detection based on feature selection and integrated learning

Zhou Chuanhua[1, 2], Liu Zhicai[1†], Ding Jing'an[1], Zhou Jiayi[3]

(1. *School of Management Science & Engineering Anhui University of Technology*, *Maanshan Anhui* 243002, *China*; 2. *School of Computer Science & Technology*, *University of Science & Technology of China*, *Hefei* 230026, *China*; 3. *Graduate School of Information*, *Production and Systems Waseda University*, *Tokyo*, *Japan*)

Abstract: In view of the fact that most phishing websites detection methods have the problems of low detection accuracy and high false positive rate and other issues, this paper proposed a phishing website detection method based on feature selection and integrated learning. Firstly, the FSIGR algorithm was used to select feature. The FSIGR algorithm combined with the advantages of filter and wrapper modes. First, it carried out a comprehensive measurement of features from two aspects of information correlation and classification ability. Second, it used recursive elimination after increasing forward strategy to select the features, and used the classification accuracy as the evaluation index to measure and select the feature subset. Finally, it obtained the optimal feature subset. Then, based on random forest integrated learning classification algorithm, it trained the selected optimal feature subset. Experiments on the UCI dataset show that this method can improve the accuracy of phishing websites detection and reduce the false positive rate.

Key Words: phishing website; random forest; information gain ratio; feature selection

0　引言

随着互联网的不断发展、普及和用户数量的增加，特别是电子商务的快速发展，互联网安全问题变的尤其重要。钓鱼网站（phishing）就是互联网安全威胁之一，它是模仿合法网站恶意创造出来的一个假网页，并使用社会工程技术对网络用户进行恶意攻击，从而获取利益和用户名、密码等私密信息[1,2]。根据反钓鱼网站工作组（APWG）的报告显示，在 2016 年第四季度，APWG 每月平均发现网络钓鱼袭击 92 564 次，与 2004 年相比 12 年间增加 5 753％；2016 年网络诈骗攻击总数为 1 220 523 次，比 2015 年增加 65％。其中，受钓鱼网站影响最严重的国家是中国，47.09％的机器受到感染[3]。

虽然攻击者使用不同的技术创建钓鱼网站来欺骗用户，但他们都使用一组常见特征来设计钓鱼网站。因此，这也给反恶意网站工作者提供了解决问题的方法与思路。目前钓鱼网站检测方法主要有用户教育[4~6]、黑名单技术[7,8]和启发式技术[9~16]等。其中，启发式技术的研究和使用较为广泛，它主要是通过提取网站的相关特征，然后应用启发式规则或者机器学习算法对特征进行处理，以达到对网页进行分类（合法/钓鱼）的目的。文献[11]通过对网页标题、关键字等进行特征提取，并采用 NBC

作者简介：周传华（1965-），男，安徽马鞍山人，教授，博士，主要研究方向为机器学习、数据挖掘、智能算法；柳智才（1993-），男（通信作者），安徽阜阳人，硕士研究生，主要研究方向为机器学习、模式识别、智能优化（lzc646211927@163.com）；丁敬安（1991-），男，安徽宿州人，硕士研究生，主要研究数据挖掘、数据分析；周家亿（1993-），男，安徽马鞍山人，硕士研究生，主要研究数据分析、智能算法、模式识别.

和 SVM 分类算法作为基分类器，采用分类集成方法综合基分类器的检测结果，提出了一种有效的钓鱼网站智能检测系统。文献[12]基于 SVM 分类算法提出了一个针对 URL 进行匹配过滤和分类识别的网络钓鱼检测系统，该系统虽然能够提高钓鱼网站预测的准确率，但仅适用于低维小样本数据。文献[13]使用 K-means 算法对 URL 特征或者页面特征进行处理，以达到预测钓鱼网站的目的。该方法虽然在一定程度上能提高预测模型的分类精度，但分类性能有限。文献[14]通过对比多元感知器、决策树和贝叶斯分类算法对钓鱼网站的预测性能发现相对于其他两种分类算法，决策树分类模型具有较优的分类性能。通过以上文献分析可知：a）通常主要从 HTML 标签、URL 地址、编码、页面图片等方面对网页进行特征提取[15,16]，特征维数较高，存在大量冗余特征，影响分类模型的准确率；b)单分类器模型分类性能有限，存在泛化能力和容错性较差等问题。

针对以上问题，本文提出一种基于特征选择和集成学习算法的钓鱼网站检测方法。其中，特征选择能有效减少大量冗余特征，从而提高钓鱼网站预测的准确率[17~20]、降低时间开销；使用集成学习算法综合各基分类器的分类结果构建分类模型，能有效提高分类模型的容错性和泛化能力，从而降低钓鱼网站预测的误判率。在特征选择阶段，本文提出了基于信息增益率和随机森林的特征选择算法（feature selection based on importance and gain rate,FSIGR）。FSIGR 特征选择算法分为过滤和封装两个阶段。在过滤阶段，以特征与类别的信息相关性为依据对特征进行选择；在封装阶段，对选择后的特征从信息相关性和分类能力两个维度计算特征权重向量和综合权重并排序，使用前项递增后向递归删除策略进行选择，并以分类精度为依据对特征子集进行评估，从而选出相关性强、冗余度低的最优特征子集，提高钓鱼网站预测的准确率。在分类阶段，使用随机森林集成学习分类算法对数据进行训练得到最终的分类模型，降低钓鱼网站预测的误判率。实验结果表明，本文提出的钓鱼网站检测方法能有效提高钓鱼网站预测的准确率，降低误判率。

1　基础理论

1.1　熵与信息增益率

香农熵作为信息论中的基本概念，是用于度量随机变量不确定性的一种数学表达，也是对变量本身或变量集合所含有的平均信息量的一种度量，通常用$H(X)$表示。设$X=\{x_1,x_2,\dots,x_m\}$与$Y=\{y_1,y_2,\dots,y_m\}$是两个随机变量，$p(x_i)$和$p(y_i)$为概率密度函数，则随机变量X的熵$H(X)$定义为

$$H(X)=-\sum_{i=1}^{m}p(x_i)\,log_2 p(x_i) \qquad (1)$$

随机变量X和Y的条件熵定义为

$$H(X|Y)=\sum_{i=1}^{m}p(y_i)\,H(X|Y=y_i) \qquad (2)$$

条件熵$H(X|Y)\le H(X)$用来衡量变量X和Y的相关性。若变量X与Y不相关，则$H(X|Y)=H(X)$；若变量X与Y相关，则$H(X|Y)<H(X)$，且$H(X)-H(X|Y)$值越大，变量X和Y相关性越强。

信息增益（information gain, IG）是对一个随机变量在另一个随机变量确定的情况下相关信息量的度量。信息增益具有非对称性，是一种无量纲的度量标准，值越大，说明变量之间的相关性越强。信息增益与熵、条件熵的关系为

$$IG(X|Y)=H(X)-H(X|Y) \qquad (3)$$

由式（3）可以看出，$IG(X|Y)$值越大，说明变量X与Y相关性越强。其中$IG(X|Y)$表示变量Y的信息增益。

在信息系统中，经常使用信息增益来衡量某个特征对信息系统分类的贡献，以降低样例中噪声的敏感度。但由于信息增益存在偏好选择分支较多的特征，导致过拟合的发生。因此，在使用时经常引入惩罚因子，以对分支较多的特征进行惩罚，即信息增益率（Gain Ratio, GR）。

$$GR(X|Y)=\frac{IG(X|Y)}{H(Y)} \qquad (4)$$

由式（4）可以看出，随机变量Y的信息增益率与其信息增益成正比，与其信息熵成反比。因此，当随机变量Y取值较多时，$GR(X|Y)$会随着$H(Y)$的增大而减小，在一定程度上降低了选择偏好的发生。

1.2　随机森林与重要度测评

随机森林（random forest, RF）是一种集成学习算法，它使用随机重采样技术和节点随机分裂技术构建多棵决策树，并根据投票机制产生最后的结果。由于 RF 对于噪声数据和存在缺失值的数据具有很好的鲁棒性，并且具有较快的学习速度，其变量重要性度量可以作为高维数据的特征选择工具，所以近年来已经被广泛应用于各种分类、预测、特征选择以及异常点检测问题中[21,22]。

基于随机森林的重要度测评，是通过袋外数据（out of bag, OOB）检测和添加随机噪声的操作来判断特征属性对输出变量的影响，影响越大，则说明该特征越重要[23~25]。

主要步骤如下：

设随机森林包括M棵分类回归树。为测度第j个特征属性对输出变量的重要性，对随机森林中的每棵分类树进行处理。对第$i(i=1,2,\dots,M)$棵分类回归树：

a）计算第i棵分类回归树基于袋外观测的预测误差率，记为e_i。

b）随机打乱袋外观测在第j个特征属性上的取值顺序，重新建立第i棵分类回归树并袋外观测进行预测。

c）重新计算第i棵分类回归树的预测误差，记为e_i^j。$\varepsilon_i^j=e_i-e_i^j$为第$j$个特征属性添加噪声导致的第i棵分类回归树预测误差的变化。

重复上述步骤，最终得到M个预测误差的变化。$\varepsilon^j=\frac{1}{M}\sum_{i=1}^{m}\varepsilon_i^j$即为第$j$个输入变量加噪声导致的随机森林总体预测误差的平均变化,它测度了第j个输入变量的重要性。

2 钓鱼网站检测模型

2.1 FSIGR 特征选择方法

特征选择是指在保证特征集合分类性能的前提下，从一组原始特征集合中选出具有代表性的特征子集，以达到降低特征空间维数的过程[26]。根据是否依赖机器学习算法，特征选择算法可以分为过滤式（filter）和封装式（wrapper）两种。过滤式特征选择算法利用数据的内在特性对选取的特征子集进行评价和选择，独立于机器学习算法，该类算法通常运行效率较高，但分类性能较差；而封装式特征选择算法则依赖于机器学习算法的分类精度作为特征子集选择的评价准则，该类算法选择的特征集合分类性能较优，但效率较低。以信息增益率和重要度测评为基础，综合过滤式和封装式特征选择算法的优点，本文提出 FSIGR 特征选择算法。

FSIGR 算法包括过滤和封装两个阶段，关键步骤如下：

a）过滤无关特征并对相关特征进行综合度量。

首先计算每个特征关于类别特征的 GR，若其 GR=0，则表示该特征和类别特征不相关，并从特征集合中删除该特征。对剩余特征子集中的每个特征计算综合度量值。

设数据集为D，特征属性集为F=$\{f_i|i=1,\dots,v\}$，首先对数据集中的特征分别使用 GR 和 RF 两种方法计算特征的信息相关性和分类能力，然后对计算结果分别进行标准化处理。具体公式如下：

$$\widetilde{m}_i=\frac{m_i}{\sum_{i=1}^{v}\mathrm{m}_i} \quad (5)$$

$$\widetilde{g}_i=\frac{g_i}{\sum_{i=1}^{v}g_i} \quad (6)$$

其中：m_i和g_i分别表示 RF 和 GR 算法对特征$f_i(i=1,\dots,v)$的权重；$\widetilde{m}_i$和$\widetilde{g}_i$则分别表示其标准化后的值。并映射成权重向量$\vec{c_i}=(\widetilde{m}_i,\widetilde{g}_i)$，其中$\widetilde{m}_i$和$\widetilde{g}_i$表示向量$\vec{c_i}$的坐标值。向量$\vec{c_i}$的长度则表示特征$f_i$的重要度。

计算特征f_i的综合评估值c_i：

$$c_i=\sqrt{\widetilde{m}_i{}^2+\widetilde{g}_i{}^2} \quad (7)$$

式（7）中，通过将$\widetilde{m}_i$和$\widetilde{g}_i$的值进行向量化并求出c_i的值对特征f_i进行综合度量，既考虑了特征f_i与类别特征之间相关性，又考虑到了特征f_i的分类能力，增强了对特征的度量，降低了特征的波动性。从而选择出最大相关和最大分类能力的特征。与文献[19]中的 IG 相比，本文使用 GR 计算特征的信息相关性降低了 IG 的选择偏好问题；与文献[23]中 MDA+MDG 的方法相比，本文从特征信息相关性和分类能力两种不同的维度对特征进行度量，并使用向量化映射关系求解特征综合度量值c_i，降低了特征的波动性。

b）采用前向递增后向递归剔除的策略进行特征选择。

根据c_i对特征进行降序排序，使用前向递增策略遍历特征空间，每次增加一个特征得到相应的特征集合$\mathrm{F}_1,\mathrm{F}_2,\dots,\mathrm{F}_v$（$v$表示特征子集的大小），并使用分类器对该特征集合进行评估,记为a_i。若$a_i<a_{i-1}$，则从集合F中删除f_i元素，直至循环结束。

该选择策略的优点是：在综合评估的基础上，使用分类精度来再次评估每个特征对整体的分类贡献，可以在不牺牲算法精度的情况下降低特征的波动性，并删除重要度较小的冗余属性。每次删除特征元素后，都会重新遍历特征集合，以产生新的特征组合，扩大特征子集搜索空间的覆盖范围，从而选出最小冗余、性能最优的特征集合。

与前向、后向搜索策略相比，本文搜索策略在排序的基础上以特征子集的整体分类性能为评价指标，递归剔除冗余且c_i最小的特征，可以在不牺牲算法精度的情况下降低特征的波动性。与文献[19,23]中的过滤式方法相比，本文采用过滤+封装的模式，提高了特征子集的分类性能。

FSIGR 算法描述如下：

输入：数据集$\boldsymbol{D}$，特征集合F=$\{f_i|i=1\dots v\}$. $\mathrm{a}_{max}=0$，$F_{best}=\emptyset$.

过程：

分别计算特征f_i关于类别特征的 GR 值g_i，若g_i=0，则删除特征f_i，$F=F-\{f_i\}$；

使用随机森林计算特征f_i重要度值，并记为m_i；

运用式（5）（6）分别对m_i，g_i进行标准化，得到$\widetilde{m}_i$，$\widetilde{g}_i$；

根据式（7）计算特征f_i的综合评估值c_i；

根据特征f_i的综合测评度c_i，对特征进行降序排序；

Repeat

使用分类器进行评估，对排序后的特征子集采用前项选择策略遍历特征空间，分别计算分类器在该特征子集F_i上的精确度a_i，其中i表示特征子集中元素的个数；

```
flag = false
for a_i（i = 1 … v）do
  if a_i < a_{i-1} then
   flag = true
   从集合 F 中删除特征f_i，并记录删除特征f_i后分类器的精度为a_temp；
    if a_max < a_temp then
      a_max = a_temp，F_best = F
    end if
    break
  end if
end for
until flag == false 达到终止条件
```

输出：最优特征子集F_{best}.

2.2 FSIGR 算法复杂度分析

算法的时间开销主要两个部分：

a)过滤阶段。根据每个特征的信息相关性对特征进行过滤，并结合特征的分类能力对特征进行综合度量。

b)封装阶段。根据特征的综合度量对特征进行排序，采用前向递增后向递归剔除搜索策略选择特征子集，并使用分类器

对特征子集进行评估。

算法时间开销主要体现在封装阶段。根据文献[22]可知，若训练数据集的特征维数为m，训练样本个数为n，假设随机森林中基分类器的个数为 k，则随机森林算法的时间复杂度近似为$O(kmn(log\,n)^2)$，快速排序平均时间复杂度为$O(m(log\,m))$。

因此，在本文算法中过滤阶段时间复杂度为$O(m+kmn(log\,n)^2)$，封装阶段外层循环最多运行m次，每次循环采用前向增加策略进行特征选择时分别进行$(m,\ m-1,\ m-2,\ldots,1)$次，采用后向递归剔除策略时，平均运行$m/2$次，最多运行$m-1$次。因此，FSIGR 算法最大时间复杂度可以近似表示为

$$O(m+kmn(log\,n)^2)+O(m(\log m))+O(m+1/2*m(m-1))+O(m-1)=O(m(1/2*(m+5)+kn(logn)^2+logm)\quad(8)$$

$$T(n)=O(m^2)\quad(9)$$

由于本文算法在运行过程中临时占用存储空间大小与特征个数成线性正比关系，所以空间复杂度可以表示为

$$S(n)=O(m)\quad(10)$$

由式(9)和(10)可以看出，FSIGR 算法的最大时间复杂度与特征维数近似平方，空间复杂度与特征维数成线性关系，因此，FSIGR 算法对高维数据具有较好的处理能力，且具有很好的扩展性。

2.3 钓鱼网站检测模型

图 1 为本文钓鱼网站检测模型。其主要包含三个部分：

a）特征提取。对网页内容进行解析，并提取相关特征；（本文实验部分数据集）。

b）特征选择。采用本文 FSIGR 特征选择算法从单个特征和特征子集两个方面对特征进行评估和选择，从而选择出相关性高，冗余度低的最优特征子集。

c）分类决策模型。使用 RF 集成学习算法构建分类决策模型，有效提高钓鱼网站检测模型的分类精度。

该模型的主要执行流程如下：首先从 HTML 标签、URL 地址、编码、页面图片等方面对网页进行特征提取，并转换成训练和预测数据；然后对提取后的特征数据使用 FSIGR 算法进行特征选择，并找出最优特征子集；最后基于选择后的最优特征子集数据对 RF 分类决策模型进行训练与结果预测。

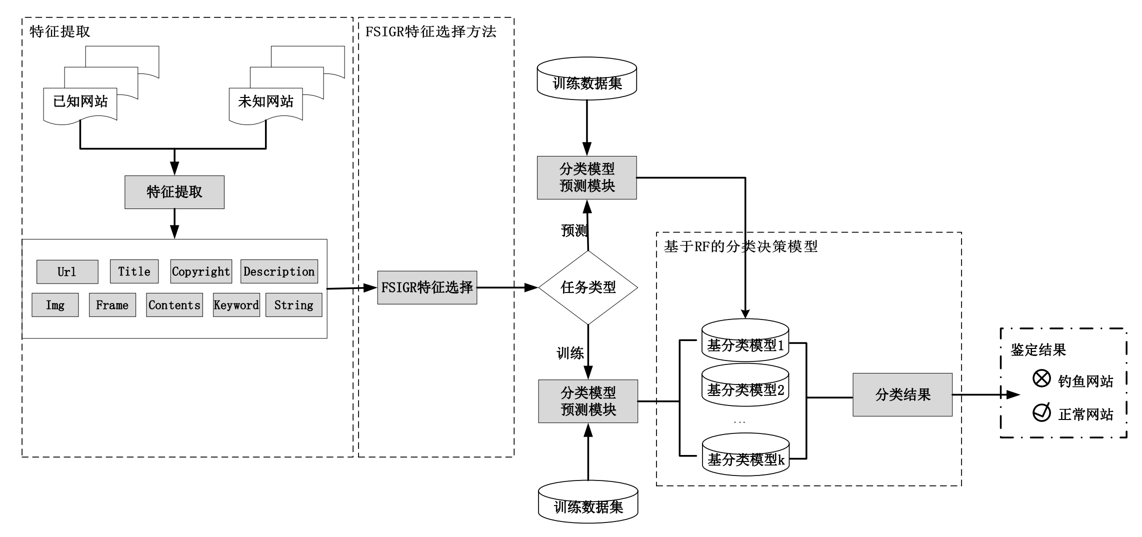


图 1 钓鱼网站检测模型

3 实验及结果分析

3.1 实验数据

本文使用 UCI 数据库中 phishing 数据集[27]进行实验。该数据集共包括 11 055 个网站实例，其中 4 898 个（44%）被标记为钓鱼网页，用-1 表示；6 157 个（56%）被标记为合法网页，用 1 表示。每个实例共包含 30 个特征，分别基于地址栏、反常标志、HTML 和 Javascript 以及域名进行提取。特征的取值是为二元（-1，1）或三元（0，1，-1）关系，更多详细信息见文献[27]。

3.2 实验说明

为了充分验证本文钓鱼网站检测方法的有效性，实验由两部分组成。

实验 1：验证 FSIGR 算法的有效性

本实验中选用 CFS（correlation-based feature selection）、WFS（wrapper feature selection）算法以及文献[19]中的算法与 FSIGR 算法进行实验对比。使用 RF 集成学习分类算法对不同特征选择算法的实验结果进行验证，并采用 10 折交叉验证的方法计算分类模型的分类精度。对比、分析实验结果，验证本文特征选择方法的有效性。

实验 2：验证本文钓鱼检测方法的有效性

在 phishing 数据集上首先使用 FSIGR 特征选择算法选出最优特征子集（以相应的分类算法作为特征子集的评估器），然后分别使用 C4.5、KNN、Naive Bayes、REP Tree 和 RF 算法进行分类模型训练，并采用 10 折交叉的方法计算分类模型的精度。对比实验结果，验证本文钓鱼检测方法的有效性。

实验软硬件环境如下：操作系统为 Windows 10，CPU 为 Intel® Core™ i5-6300HQ @ 2.3 GHz，实验内存为 8 GB，主要实

验平台为 WEKA，语言为 Java。

3.3 评判指标

一般采用精确度和召回率两个指标对分类算法性能进行评判。

1）精确度（*accuracy*）：又叫查准率，计算公式为

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{11}$$

2）召回率（recall）：又叫查全率，计算公式为

$$recall = \frac{TP}{TP+FN} \tag{12}$$

其中：TP（true positive）为被正确分类为正例的样本数；FP（false positive）为被错误分类为正例的样本数；TN（true negative）为被正确分类为反例的样本数；FN（false negative）为被错误分类为反例的样本数。

3.4 结果分析

实验中，CFS 和 WFS 算法分别采用 BF(best first)和 GS(greedy stepwise)两种搜索算法对特征进行选择；FSIGR 算法则采用前向递增后向递归剔除策略对特征进行选择。具体实验结果如下：

表 2 中列出了在 phishing 数据集上使用不同特征选择算法进行特征选择，并使用 RF 集成学习分类算法对特征子集进行训练，10 折交叉验证的实验结果。其中，根据文献[19]以阈值 0.01 对 GR 和 RF 重要度排序后的特征进行选择，SF 表示特征的个数，Acc 表示分类精度，M-error 表示平均绝对误差。AUC 表示 ROC 曲线的面积。

表 2 基于不同特征选择算法构建 RF 分类预测模型实验结果/%

特征选择算法	SF	Acc	recall	M-error	AUC
GR	11	95.215±3.448	95.2	6.50	99.1
RF	13	96.002±2.769	96.0	5.44	99.3
WFS(BF)	28	97.205±2.056	97.2	5.01	99.6
WFS(GS)	29	97.286±2.048	97.3	5.09	99.6
CFS(BF/GS)	9	94.772±3.873	94.8	7.46	98.8
文献[21]算法	17	96.834±2.292	96.8	4.87	99.5
FSIGR	23	**97.341±2.053**	**97.3**	**4.80**	**99.6**

注：表中±前面和后面的数据分别表示 10 次测试结果的平均分类精度和方差。

由表 2 可以看出，本文 FSIGR 特征选择方法分类精度为 97.341%，召回率为 97.3%，平均绝对误差为 0.048，均优于其他特征选择方法。其中，文献[21]算法的分类精度为 96.834%，召回率为 96.8%，平均绝对误差为 0.048 7，分类模型性能明显低于 FSIGR 方法的分类模型。CFS、GR 和 RF 特征选择方法在特征降维方面表现较优，选择后的特征子集大小分别为 9、11 和 13，但分类精度较低。WFS 特征选择方法两种搜索策略选择的特征子集大小分别为 28 和 29，在特征降维方面性能低于其他方法，分类精度分别为 97.205%和 97.286%，优于 CFS 等方法，但与本文方法相比综合性能较差且时间代价较大。实验结果表明，本文 FSIGR 特征选择方法能够选出特征维度较低，分类性能最优的特征子集，满足实际应用需求，证明了其方法的有效性。

表 3 中列出了在 phishing 数据集上基于 FSIGR 特征选择方法使用不同分类算法与本文方法实验结果对比。实验结果均采用 10 折交叉验证产生。

表 3 不同分类算法基于 FSIGR 特征选择算法构建分类预测模型实验结果/%

分类算法	SF	Acc	recall	M-error	AUC
C4.5	25	96.056±3.312	96.1	5.68	98.5
KNN	25	97.205±2.091	97.2	**3.28**	99.0
REP Tree	28	95.432±3.602	95.4	6.39	98.5
Naive Bayes	28	92.999±5.308	93.0	8.94	98.1
RF	**23**	**97.341±2.053**	**97.3**	4.80	**99.6**

注：表中±前面和后面的数据分别表示 10 次测试结果的平均分类精度和方差。

由表 3 可知，本文方法的分类精度为 97.341%，分类召回率为 97.3%，平均绝对误差为 0.048，特征子集维数为 23，综合分类性能明显优于 C4.5、REPTree 和 NaiveBayes 算法。与 KNN 算法相比，虽然平均绝对误差高于 KNN 的 0.328，但其综合性能优于 KNN 算法。由实验结果可知，本文提出的钓鱼网站检测方法分类性能明显优于 C4.5、KNN、REPTree、NaiveBayes 算法的分类性能，验证了本文方法的有效性。

受试者工作特征(receiver operating characteristic, ROC)曲线体现了综合考虑分类模型在不同任务下的泛化性能，ROC 曲线下的面积，即 AUC（area under ROC curve）越大，则表示该分类模型的泛化能力越强。由表 2 可以看出，在同种分类器下本文提出的 FSIGR 算法的 AUC 值为 0.996，优于其他特征选择算法，证明了本文 FSIGR 算法的适用性。由表 3 可以看出，在同种特征选择算法下，RF 分类模型的 AUC 值为 0.996，优于其他分类模型，证明了 RF 集成学习模型具有较强的容错性。因此，本文提出的钓鱼网站检测模型具有较强的泛化能力。

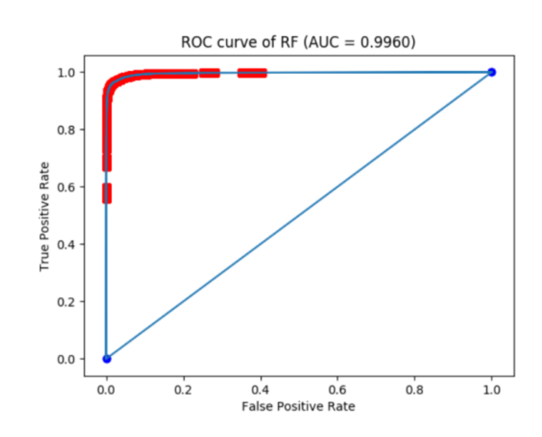


图 2 基于 FSIGR 算法 RF 分类决策模型 ROC 曲线图

图 3、4 中描述了 C4.5 算法在 phishing 数据集和最优特征子集上不同特征维度的分类精度变化折线图。图 3 中，蓝色三

角代表特征子集中加入当前特征后分类精度不变，红色三角代表特征子集中加入当前特征后分类精度下降（见电子版）。图 5 中描述了 RF 集成学习算法在最优特征子集上不同特征维度的分类精度变化折线图。

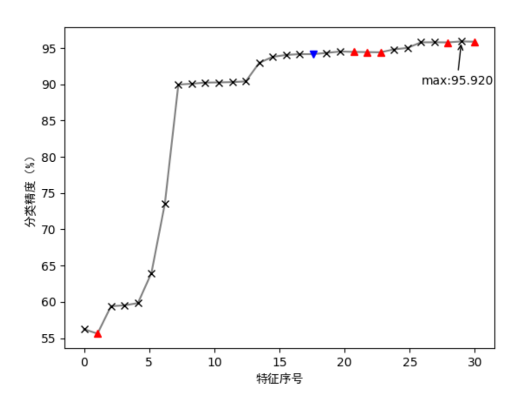


图 3　在 phishing 数据集上不同特征维度 C4.5 分类精度变化折线图

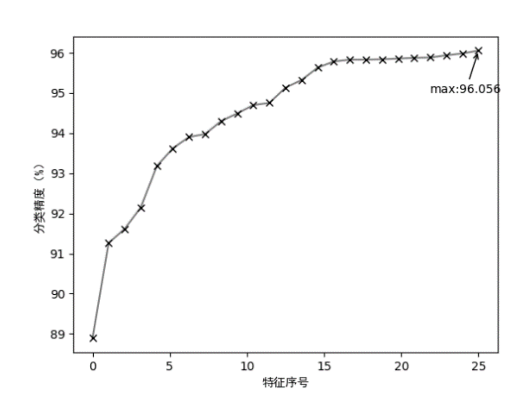


图 4　在 FSIGR 选择的最优特征子集上不同特征维度 C4.5 分类精度折线图

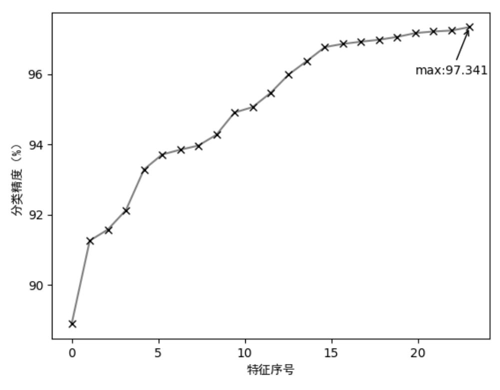


图 5　在 FSIGR 选择的最优特征子集上不同特征维度 RF 分类精度折线图

由图 3、4 对比可以看出，在原始数据中随着特征个数的增加会出现分类精度不变（图 3 中的特征 18）甚至下降（图 3 中的特征 2、21、22、23、28 和 30）的情况。而在图 4 中，随着特征维数的增加分类精度则不断提高，且最大分类精度为 96.056%，特征维度为 25，优于原数据集的 95.920%和 30。时间开销从 0.17 s 降低到 0.1 s。这是因为在原数据集中存在部分无关特征和冗余特征，导致了分类器分类性能的下降，使用 FSIGR 方法能够从信息相关性和分类能力两个方面对特征进行综合度量，从而选出相关性强、冗余度低的最优特征子集，提高了分类器的分类精度。本实验证明了 FSIGR 特征算法能有效降低特征子集的维度选出关键特征，从而提高分类模型的准确率。

由图 4、5 可以看出，RF 集成学习算法分类精度为 97.341%，特征维度为 23 明显优于 C4.5 单分类器的 96.056%和 25。这是因为集成学习算法能够通过综合不同基分类器模型的分类结果增强集成学习算法的容错性和泛化能力，从而达到提高分类精度，分类召回率降低分类误差的目的。本文实验证通过明了集成学习算法对钓鱼网站检测的有效性，从而证明力本文钓鱼网站检测方法的有效性。

4　结束语

本文提出来了一种基于特征选择和集成学习的钓鱼网站检测方法。该方法首先运用 FSIGR 算法选择出相关性强、冗余度低的最优特征子集，然后使用最优特征子集数据集基于 RF 集成学习分类算法进行分类模型训练来提高分类预测模型的准确率。通过 FSIGR 算法和 CFS、WFS 以及文献[19]算法在 phishing 数据集上的实验结果表明 FSIGR 算法在特征降维和提高分类精度方面均有很好的表现，证明了 FSIGR 算法的有效性。通过对 FSIGR 算法进行时间复杂都分析发现，FSIGR 算法对高维数据有较好的处理能力，具有较好的扩展性。通过 RF 集成学习算法和 C4.5、KNN、REPTree 以及 NaiveBayes 算法在 phishing 数据集上的实验表明 RF 集成学习算法分能性能明显优于其他单分类器模型，具有分类准确率高、分类误差率低和召回率高等优点。基于以上叙述，证明了本文钓鱼网站检测方法的有效性和实际应用性。

利用关联信息熵对组合特征相关性进行排序，选择最优特征子集，构建钓鱼网站的检测模型，提高模型预测的准确率是笔者下一步工作的重点。

参考文献：

[1] Almomani A, Gupta B B, Atawneh S, et al. A survey of phishing email filtering techniques [J]. IEEE Communications Surveys & Tutorials, 2013, 15 (4): 2070-2090.

[2] Mishra A, Gupta B B. Hybrid solution to detect and filter zero-day phishing attacks [C]// Proc of the 2nd International Conference on Emerging Research in Computing, Information, Communication and Applications. 2014: 373-379.

[3] Anti-Phishing Working Group. Phishing activity trends report of 4th quarter of 2016 [R]. 2016.

[4] Sheng S, Holbrook M, Kumaraguru P, et al. Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions [C]// Proc of Sigchi Conference on Human Factors in Computing Systems. 2010: 373-382.

[5] Arachchilage N A G, Love S. A game design framework for avoiding phishing attacks [J]. Computers in Human Behavior, 2013, 29 (3): 706-714.

[6] Zhuang W, Jiang Q. Intelligent anti-phishing framework using multiple classifiers combination [J]. Journal of Computational Information Systems, 2012, 8 (17): 7267-7281.

[7] Zhang J, Porras P, Ullrich J. Highly predictive blacklisting [C]// Proc of Conference on Security Symposium. USENIX Association. 2008: 107-122.

[8] Sharifi M, Siadati S H. A phishing sites blacklist generator [C]// Proc of IEEE//ACS International Conference on Computer Systems and Applications. 2008: 840-843.

[9] Zhang Y, Hong J I, Cranor L F. CANTINA: a content-based approach to detecting phishing web sites [C]// Proc of International Conference on World Wide Web. 2007: 639-648.

[10] Xiang G, Hong J, Rose C P, et al. CANTINA+: a feature-rich machine learning framework for detecting phishing web sites [J]. ACM Trans on Information & System Security, 2011, 14 (2): 21

[11] 庄蔚蔚，叶艳芳，李涛，等. 基于分类集成的钓鱼网站智能检测系统 [J]. 系统工程理论与实践, 2011, 31 (10): 2008-2020.

[12] 何高辉，邹福泰，谭大礼，等. 基于 SVM 主动学习算法的网络钓鱼检测系统 [J]. 计算机工程, 2011, 37 (19): 126-128.

[13] Sahu K K. Shrivastava S. Kernel K-means clustering for phishing website and malware categorization [J]. International Journal of Computer Applications, 2015, 111 (9): 20-25.

[14] Lakshmi V S, Vijaya M S. Efficient prediction of phishing websites using supervised learning algorithms [J]. Procedia Engineering, 2012, 30 (9): 798-805.

[15] Pan Y, Ding X. Anomaly based Web phishing page detection [C]// Proc of the 22nd Computer Security Applications Conference. 2006: 381-392.

[16] Basnet R B, Sung A H, Liu Q. Rule-based phishing attack detection [C]// Proc of International Conference on Security and Management. 2011.

[17] Zuhair H, Selmat A, Salleh M. The effect of feature selection on phish website detection [J]. International Journal of Advanced Computer Science & Applications, 2015, 6 (10): 221-232.

[18] Zhang W, Ren H, Jiang Q. Application of feature engineering for phishing detection [J]. IEICE Trans on Information & Systems, 2016, 99 (4): 1062-1070.

[19] Rajab K D. New hybrid features selection method: a case study on websites phishing [J]. Security & Communication Networks, 2017, 2017 (2): 1-10.

[20] Basnet R B, Sung A H, Liu Q. Feature selection for improved phishing detection [C]// Proc of the 25th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems. 2012, 7345: 252-261.

[21] Hideko K, Hiroaki Y. Rapid feature selection based on random forests for high-dimensional data [J]. Ipsj Sig Notes, 2012, 2012: 1-7.

[22] 姚登举，杨静，詹晓娟. 基于随机森林的特征选择算法 [J]. 吉林大学学报：工学版, 2014, 44 (1): 137-141.

[23] Han H, Guo X, Yu H. Variable selection using mean decrease accuracy and mean decrease gini based on random forest [C]// Proc of IEEE International Conference on Software Engineering and Service Science. 2017: 219-224.

[24] Wang H, Lin C, Peng Y, et al. Application of improved random forest variables importance measure to traditional Chinese chronic gastritis diagnosis [C]// Proc of IEEE International Symposium on It in Medicine and Education. 2008: 84-89.

[25] Nicodemus K K. Letter to the editor: on the stability and ranking of predictors from random forest variable importance measures [J]. Briefings in Bioinformatics, 2011, 12 (4): 369-373.

[26] Guyon, Isabelle, Elisseeff, et al. An introduction to variable and feature selection [J]. Journal of Machine Learning Research, 2003, 3 (6): 1157-1182.

[27] Mohammad R M, Thabtah F, McCluskey L. Phishing websites features [EB/OL]. (2015) . http://eprints. hud. ac. uk/24330/6/RamiPhishing_Websites_Features. pdf.